IN THE CLAIMS:

Please substitute the following claims for the same-numbered claims in the application:

1. (Original) A method for determining a degree of similarity between documents in a given document collection, the method comprising ~~the steps of~~:

modeling all said documents as labeled tree representations;

building a computerized dictionary of path representations relating to paths that occur in said documents;

storing, for at least two said documents, said labeled tree representations of respective documents;

storing, for said at least two said documents, said path representations relating to said paths that occur in ~~the~~ said documents from root nodes to leaf nodes in ~~the~~ said labeled tree representations of ~~the~~ said respective documents; ~~and~~

representing each of said documents in said document collection as an *N*-dimensional vector comprising an element *i* denoting a value of a feature associated with a particular path, wherein said feature comprises any of a presence or absence of said particular path in said documents and a frequency of occurrence of said particular path in said documents;

calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations[[.]]; and

using said measure of similarity to cluster a plurality of documents comprising similar information, wherein said documents comprise any of web page documents and eXtensible Markup Language (XML) documents.

wherein two documents that differ only in the frequency of occurrence of the paths associated with said two documents are considered to be more similar to each other than two documents that differ in the occurrence of paths.

2. (Currently Amended) The method as claimed in claim 1, wherein the tree representation is a Document ~~Model~~ Object Model representation.

3. (Original) The method as claimed in claim 1, further comprising the step of generating a path representation for a path of a document as a sequence of labels representative from a root node to a leaf node in the labeled tree representation of the document.

4. (Original) The method as claimed in claim 1, further comprising the step of storing, as path representations, sets of sequenced labels representative of distinct paths in a labeled tree representation of a corresponding document.

5. (Original) The method as claimed in claim 4, further comprising the step of storing a path dictionary ($Dict_{paths} = \{p_1, p_2, ..., p_N\}$) of distinct paths collated from a tree representation for a document.

6. (Original) The method as claimed in claim 5, further comprising the step of eliminating selected paths from the path dictionary ($Dict_{paths}$).

7. (Original) The method as claimed in claim 6, wherein paths that occur highly frequently or highly infrequently are eliminated from the path dictionary ($Dict_{paths}$).

8. (Original) The method as claimed in claim 7, further comprising the step of computing the frequency of occurrence ($f_j(p_i)$) of a path (p_i) in a document (d_j).

9. (Original) The method as claimed in claim 8, further comprising the step of computing the maximum number of instances ($f_{max} = \max_{ij} f_j(p_i)$) in which a path ($p_i$) in the document ($d_j$) occurs.

10. (Original) The method as claimed in claim 9, further comprising the step of storing a representation of the document (d_j) as a N-dimensional vector ($[d_{j1}, d_{j2}, .., d_{jN}]$, where $d_{jk} = f_j(p_k)/f_{max}$, $1 \leq k \leq N$) of relative frequencies of occurrence ($f_j(p_k)$) of paths (p_k) in the document (d_j).

11. (Original) The method as claimed in claim 8, further comprising the step of computing the minimum number of instances ($f_{min} = \min_{ij} f_j(p_i)$) in which a path ($p_i$) in the document ($d_j$) occurs.

12. (Original) The method as claimed in claim 10, further comprising the step of computing the similarity between a pair of documents (d_i, d_l) as a function ($sim(d_i, d_l)$) of metrics relating the number of paths common to the respective documents (d_i , d_l).

13. (Original) The method as claimed in claim 12, wherein the function for computing the similarity between a pair of documents (d_i , d_l)

$$(sim(d_i, d_l) = sim(d_l, d_l) = \frac{\sum_{k=1}^{N} \min(d_{ik}, d_{lk})}{\sum_{k=1}^{N} \max(d_{ik}, d_{lk})})$$

is the quotient of a numerator, defined as the sum for all paths ($k = 1 \ldots N$) of the minimum number of instances ($\min(d_{ik}, d_{lk})$) in which paths occur in the respective documents (d_i, d_l), and a denominator, defined as the sum for all paths ($k = 1 \ldots N$) of the maximum number of instances ($\min(d_{ik}, d_{lk})$) in which paths occur in the respective documents (d_i, d_l).

14. (Original) The method as claimed in claim 1, wherein the tree representation of a document includes a positional index, which represents, for a node (n), the number of previous sibling nodes with the same label as that of node (n).

15. (Original) The method as claimed in claim 14, further comprising the step of storing as a path representation a set that defines positional information of sibling nodes under a parent node.

16. (Original) The method as claimed in claim 15, further comprising the step of storing precise path representations that precisely define a document structure, and generalised path

representations that partially generalise structural aspects of precise path representations of a document.

17. (Original) The method as claimed in claim 16, wherein the step of calculating the measure of similarity involves determining a total number of precise path representations of one document that are either shared by the other document, or are a subsumed subset of at least one of the generalised path representations of the other document.

18. (Original) The method as claimed in claim 17, further comprising the step of normalising the measure of similarity by a term that represents the number of unique path representations shared by the two documents.

19. (Original) The method as claimed in claim 18, wherein the number of unique path representations is calculated by adding the number of path representations for each document, and subtracting from this total the number path representations shared by the two documents.

20. (Original) The method as claimed in claim 14, further comprising the step of storing as a path representation a sequence of terms separated by a delimiting symbol, in which each term is represented by a label and a parenthesised predicate that specifies the positional index of the term either specifically or generally.

21-22. (Cancelled).

23. (Currently Amended) A program storage device readable by computer, tangibly embodying a program of instructions executable by said computer to perform a method for determining a degree of similarity between documents in a given document collection, the method comprising:

modeling all said documents as labeled tree representations;

building a computerized dictionary of path representations relating to paths that occur in said documents;

storing, for at least two said documents, said labeled tree representations of respective documents;

storing, for said at least two said documents, said path representations relating to said paths that occur in ~~the~~ said documents from root nodes to leaf nodes in ~~the~~ said labeled tree representations of ~~the~~ said respective documents; ~~and~~

representing each of said documents in said document collection as an *N*-dimensional vector comprising an element *i* denoting a value of a feature associated with a particular path, wherein said feature comprises any of a presence or absence of said particular path in said documents and a frequency of occurrence of said particular path in said documents;

calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations[[.]]; and

using said measure of similarity to cluster a plurality of documents comprising similar information, wherein said documents comprise any of web page documents and eXtensible Markup Language (XML) documents.

wherein two documents that differ only in the frequency of occurrence of the paths associated with said two documents are considered to be more similar to each other than two documents that differ in the occurrence of paths.

24. (Currently Amended) The program storage device in claim 23, wherein said method further comprises the tree representation is a Document ~~Model~~ Object Model representation.

25. (Previously Presented) The program storage device in claim 23, wherein said method further comprises the step of generating a path representation for a path of a document as a sequence of labels representative from a root node to a leaf node in the labeled tree representation of the document.

26. (Previously Presented) The program storage device in claim 23, wherein said method further comprises the step of storing, as path representations, sets of sequenced labels representative of distinct paths in a labeled tree representation of a corresponding document.

[[28]] 27. (Currently Amended) The program storage device in claim 23, wherein the tree representation of a document includes a positional index, which represents, for a node (n), the number of previous sibling nodes with the same label as that of node (n).

[[29]] 28. (Currently Amended) A computer system operable for determining a degree of similarity between documents in a given document collection, the computer system comprising:

~~a first storage unit operable for storing labeled tree representations of respective documents for at least two documents;~~

~~a second storage unit operable for storing, for at least two documents, path representations relating to paths that occur in the documents from root nodes to leaf nodes in the labeled tree representations of the respective documents; and~~

~~a calculator operable for calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations.~~

means for modeling all said documents as labeled tree representations;

means for building a computerized dictionary of path representations relating to paths that occur in said documents;

means for storing, for at least two said documents, said labeled tree representations of respective documents;

means for storing, for said at least two said documents, said path representations relating to said paths that occur in said documents from root nodes to leaf nodes in said labeled tree representations of said respective documents;

means for representing each of said documents in said document collection as an *N*-dimensional vector comprising an element *i* denoting a value of a feature associated with a particular path, wherein said feature comprises any of a presence or absence of said particular path in said documents and a frequency of occurrence of said particular path in said documents;

means for calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations; and

means for using said measure of similarity to cluster a plurality of documents comprising similar information, wherein said documents comprise any of web page documents and eXtensible Markup Language (XML) documents,

wherein two documents that differ only in the frequency of occurrence of the paths associated with said two documents are considered to be more similar to each other than two documents that differ in the occurrence of paths.

[[30]] 29. (Currently Amended) The computer system device in claim [[29]] 28, wherein said ~~method further comprises the~~ tree ~~representation~~ representations is a Document ~~Model~~ Object Model representation.

[[31]] 30. (Currently Amended) The computer system device in claim [[29]] 28, ~~wherein said method further comprises the step of~~ further comprising means for generating a path representation for a path of a document as a sequence of labels representative from a root node to a leaf node in the labeled tree representation of the document.

[[32]] 31. (Currently Amended) The computer system device in claim [[29]] 28, ~~wherein said method further comprises the step of~~ further comprising means for storing, as path representations, sets of sequenced labels representative of distinct paths in a labeled tree representation of a corresponding document.

[[33]] 32. (Currently Amended) The computer system device in claim [[29]] 28, wherein the tree representation of a document includes a positional index, which represents, for a node (*n*), the number of previous sibling nodes with the same label as that of node (*n*).